

Optimizing Static and Adaptive Probing Schedules for Rapid Event Detection

Ahmad Mahmoody, Evgenios M. Kornaropoulos, and Eli Upfal

Department of Computer Science, Brown University
 {ahmad, evgenios, eli}@cs.brown.edu

Abstract. We formulate and study a fundamental search and detection problem, *Schedule Optimization*, motivated by a variety of real-world applications, ranging from monitoring content changes on the web, social networks, and user activities to detecting failure on large systems with many individual machines.

We consider a large system consists of many nodes, where each node has its own rate of generating new events, or items. A monitoring application can probe a small number of nodes at each step, and our goal is to compute a probing schedule that minimizes the expected number of undiscovered items at the system, or equivalently, minimizes the expected time to discover a new item in the system.

We study the Schedule Optimization problem both for deterministic and randomized memoryless algorithms. We provide lower bounds on the cost of an optimal schedule and construct close to optimal schedules with rigorous mathematical guarantees. Finally, we present an adaptive algorithm that starts with no prior information on the system and converges to the optimal memoryless algorithms by adapting to observed data.

1 Introduction

We introduce and study a fundamental stochastic search and detection problem, *Schedule Optimization*, that captures a variety of practical applications, ranging from monitoring content changes on the web, social networks, and user activities to detecting failure on large systems with many individual machines.

Our optimization problem consists of a large set of units, or *nodes*, that generate events, or *items*, according to a random process with known or unknown parameters. A detection algorithm can discover new items in the system by probing a small number of nodes in each step. This setting defines a discrete, infinite time process, and the goal of the stochastic optimization problem is to construct a probing schedule that minimizes the long term expected number of undiscovered items in the system, or equivalently, minimizes the expected time to discover a new item in the system.

We outline several important applications of this schedule optimization problem:

News and Feed Aggregators. To provide up to date summary of the news, news aggregator sites need to constantly browse the Web, and often also the blogosphere and social networks, for new items. Scanning a site for new items requires significant communication and computation resources, thus the news aggregator can scan only a few sites simultaneously. The frequency of visiting a site has to depend on the likelihood of finding new items in that site. [9,21,1]

Algorithmic Trading on Data. An emerging trend in algorithmic stock trading is the use of automatic search through the Web, the blogosphere, and social networks for relevant information that can be used in fast trading, before it appears in the more popular news sites [4,14,5,17,11,8,16]. The critical issue in this application is the speed of discovering new events, but again there is a resource limit on the number of sites that the search algorithm can scan simultaneously.

Detecting Anomaly and Machine Malfunction. In large server farm or any other large collection of semi-autonomous machines a central controller needs to identify and contain anomalies and malefactions as soon as possible, before they spread in the system. To minimize interference with the system's operation the controller must probe only a small number of machines in each step.

1.1 Our Contribution

We consider an infinite, discrete time process in which n nodes generate new items according to a stochastic process which is governed by a generating vector π (see Section 3 for details). An algorithm can probe up to c nodes per step to discover all new items in these nodes. The goal is to minimize the *cost* of the algorithm (or the probing schedule), which we define as the long term (steady state) expected number of undiscovered items in the system.

We first show that the obvious approach of probing at each step the nodes with maximum expected number of undiscovered items at that step is not optimal. In fact, the cost of such a schedule can be arbitrary far from the optimal.

Our first result toward the study of efficient schedules is a lower bound on the cost of any deterministic or random schedule as a function of the generating vector π .

Next we assume that the generating vector π is known and study explicit constructions of deterministic and random schedules. We construct a deterministic schedule whose cost is within a factor of $(3 + (c - 1)/c)$ of the optimal cost, and a very simple, memoryless random schedule with cost that is within a factor of $(2 + (c - 1)/c)$ from optimal, where c is the maximum number of probes at each step.

Finally, we address the more realistic scenario in which the generating vector, π , is not known to the algorithm and may change in time. We construct an adaptive scheduling algorithm that learns from probing the nodes and converges to the optimal memoryless random schedule.

2 Related Work

The *News and Feed Aggregation* problem is a very well-studied topic, in which the general goal is to obtain the updates of news websites (e.g. by RSS feeds). Among many introduced objectives [19,9,1] in studying this problem, the most similar one to our cost function is the *delay* function presented by [21]. In [21] it is assumed that the rates of the news publication does not change, where in our setting these rates may change and our algorithm (**Adaptive**) can adapt itself to the new setting. Also, we assume at any given time the number of probes is fixed (or bounded) regarding the limited computational power for simultaneous probes, but [21] uses a relaxed assumption by fixing the number of probes over a *time window* of a fixed length which may result in high number of probes at a single time step. Finally, [21] introduces a deterministic algorithm in which the number of probes to each feed is obtained by applying the Lagrange multipliers method (very similar result to Theorem 3), but they loose the guarantee on optimality of their solution, by rounding the estimated number of probes to integers. In contrast, our solution provides theoretical guarantee on optimality of our output schedule.

Web-crawling is another related topic, where a web-crawler aims to obtain the most recent snapshots of the web. However, it differs from our model substantially: in web-crawling algorithm data get *updated*, so missing some intermediate snapshot would not affect the quality of the algorithm, where in our model data are generated and they all need to be processed [3,22].

There has been an extensive work on *Outbreak Detection* (motivated in part by the “Battle of Water Sensors Network” challenge [20]) using statistic or mobile sensor in physical domains, and regarding a variety of objectives [13,10,7]. Our model deviates from the Outbreak Detection problem as it is geared to detection in virtual networks such as the Web or social networks embedded in the Internet, where a monitor can reach (almost) any node at about the same cost.

Another related problem is the *Emerging Topic Detection* problem, where the goal is to identify emergent topics in a social network, assuming full access to the stream of all postings. Besides having different objectives, our model differs mainly in this accessibility assumption: the social network providers have an immediate access to all tweets or postings as they are submitted to their servers, whereas in our model we consider an outside observer who needs an efficient mechanism to monitor changes, without having such full access privilege [2,15].

In the next section, we formally define our model and the Schedule Optimization problem.

3 Model and Problem Definition

We study an infinite, discrete time process in which a set of n *nodes*, indexed by $1, \dots, n$, generate new *items* according to a random generating process. The generating process at a given time step is characterized by a *generating vector* $\pi = (\pi_1, \dots, \pi_n)$, where π_i is the expected number of new items generated at

node i at that step (by either a Bernoulli or a Poisson process). The generation processes in different nodes are independent.

We focus first on a *static generating process* in which the generating vector does not change in time. We then extend our results to adapt to generating vectors that change in time.

Our goal is to detect new events as fast as possible by probing in each step a small number of nodes. In particular, we consider probing schedules that can probe up to c nodes per step.

Definition 1 (Schedule). A c -schedule is a function $\mathcal{S} : \mathbb{N} \rightarrow \{1, \dots, n\}^c$ specifying a set of c nodes to be probed at any time $t \in \mathbb{N}$. A deterministic function \mathcal{S} defines a deterministic schedule, otherwise the schedule is random.

Definition 2 (Memoryless Schedule). A random schedule is memoryless if it is defined by a vector $p = (p_1, \dots, p_n)$ such that at any step the schedule probes a set C of c items with probability $\prod_{j \in C} p_j$ independent of any other event. In that case we use the notation $\mathcal{S} = p$.

Definition 3 (Cyclic Schedule). A schedule, \mathcal{S} , is ℓ -cyclic if there is a finite time t_0 such that from time t_0 on, the schedule repeats itself every period of ℓ steps. A schedule is cyclic if it is ℓ -cyclic for some positive integer ℓ .

The quality of a probing schedule is measured by the speed in which it discovers new items in the system. When a schedule probes a node i at a time t , all items that were generated at that node by time $t - 1$ are discovered (thus, each item is not discovered in at least one step). We define the *cost* of a probing schedule as the long term expected number of undiscovered items in the system.

Definition 4 (Cost). The cost of schedule \mathcal{S} in a system of n nodes with generating vector π is

$$\text{cost}(\mathcal{S}, \pi) = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{t'=1}^t \mathbb{E}[Q^{\mathcal{S}}(t')] = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{t'=1}^t \sum_{i=1}^n \mathbb{E}[Q_i^{\mathcal{S}}(t')],$$

where $Q_i^{\mathcal{S}}(t')$ is the number of undiscovered items at node i and at time t' , and $Q^{\mathcal{S}}(t') = \sum_{i=1}^n Q_i^{\mathcal{S}}(t')$. The expectation is taken over the distribution of the generating system and the probing schedule.

While the cost can be unbounded for some schedules, the cost of the optimal schedule is always bounded. To see that, consider a round-robin schedule, \mathcal{S} , that probes each node every n steps. Clearly no item is undiscovered in this schedule for more than n steps, and the expected number of items generated in an interval of n steps is $n \sum_{i=1}^n \pi_i$. Thus, $Q^{\mathcal{S}}(t) \leq n \sum_{i=1}^n \pi_i$, which implies $\text{cost}(\mathcal{S}, \pi) \leq n \sum_{i=1}^n \pi_i$. Therefore, without loss of generality we can restrict our discussion to bounded cost schedules. Also, note that when the sequence $\{\mathbb{E}[Q^{\mathcal{S}}(t)]\}_{t \in \mathbb{N}}$ converges we have $\text{cost}(\mathcal{S}, \pi) = \lim_{t \rightarrow \infty} \mathbb{E}[Q^{\mathcal{S}}(t)]$ (Cesaro Means [6]).

One can equivalently define the cost of a schedule in terms of the expected time that an item is in the system until it is discovered.

Lemma 1. Let ω_i^S be the expected waiting time of an item generated at node i until node i is probed by schedule \mathcal{S} . Then

$$\text{cost}(\mathcal{S}, \pi) = \sum_{i=1}^n \pi_i \omega_i^S.$$

Proof. Following the definition of the cost function we have

$$\text{cost}(\mathcal{S}, \pi) = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{t'=1}^t \sum_{i=1}^n \mathbb{E}[Q_i^S(t')] = \sum_{i=1}^n \left[\lim_{t \rightarrow \infty} \frac{\sum_{t'=1}^t \mathbb{E}[Q_i^S(t')]}{t} \right] = \sum_{i=1}^n \pi_i \omega_i^S,$$

where the last equality is obtained by applying Little's Law [12]. \square

Corollary 1. A schedule that minimizes the expected number of undiscovered items in the system simultaneously minimizes the expected time that an item is undiscovered.

Corollary 2. For any schedule \mathcal{S} , $\text{cost}(\mathcal{S}, \pi) \geq \sum_{i=1}^n \pi_i$.

Proof. As mentioned above, when we probe a node i at time t we discover only the items that have been generated by time $t - 1$. Therefore, $\omega_i^S \geq 1$, and by Lemma 1 the proof is complete. \square

Now, our main problem is defined as the following:

Definition 5 (Schedule Optimization). Given a generating vector π and a positive integer c , find a c -schedule with minimum cost.

When the generating vector is not known a priori to the algorithm the goal is to design a schedule that *converges* to an optimal one. For that we need the following definition:

Definition 6 (Convergence). We say schedule \mathcal{S} converges to schedule \mathcal{S}' , if for any generating vector π , $\lim_{t \rightarrow \infty} \left| \mathbb{E}[Q^S(t)] - \mathbb{E}[Q^{S'}(t)] \right| = 0$.

4 Results

We start this section by, first, showing that the obvious approach of maximizing the expected number of detections at each step is far from optimal. We then prove a lower bound on the cost of any schedule, and provide deterministic and memoryless c -schedules that are within a factor of $(3 + (c - 1)/c)$ and $(2 + (c - 1)/c)$, respectively, from the optimal. Finally, we introduce an algorithm, **Adaptive**, which outputs a schedule \mathcal{A} that converges to the optimal *memoryless* 1-schedule when the generating vector π is not known in advance. We also show that **Adaptive** can be used to obtain a c -schedule \mathcal{A}^c whose cost is within $(2 + (c - 1)/c)$ factor of any optimal c -schedule.

Throughout this section, by $\tau_i^S(t)$ we mean the number of steps from the last time that node i was probed until time t , while executing schedule \mathcal{S} ; if i has

not been probed so far, we let $\tau_i^S(t) = t$. Using the definition, it is easy to see that

$$\mathbb{E}[Q_i^S(t)] = \pi_i \mathbb{E}[\tau_i^S(t)], \quad (1)$$

when the expectations are over the randomness of *both* \mathcal{S} and π . Therefore, if the expectation is over *only* the randomness of π we have

$$\mathbb{E}[Q_i^S(t)] = \pi_i \tau_i^S(t). \quad (2)$$

4.1 On Maximizing Immediate Gain

Let \mathcal{S} be a 1-schedule that at each step, probes the node with the maximum expected number of undetected items. By (2), the expected number of undetected items at node i and at time t is $\pi_i \tau_i^S(t)$, and thus, $\mathcal{S}(t) = \arg \max_i \pi_i \tau_i^S(t)$.

Now, suppose $\pi_i = 2^{-i}$, for $1 \leq i \leq n$. Since the probability that node 1 has an undetected item in each step is at least $1/2$, node i is probed no more than once in each 2^{i-1} steps. Thus, the expected number of time steps that an item at node i will stay undetected is at least $\frac{1}{2^{i-1}}(1 + \dots + 2^{i-1}) = \frac{2^{i-1}+1}{2} > 2^{i-2}$. Using Lemma 1, the cost of this schedule is at least $\sum_{i=1}^n \pi_i \omega_i > \sum_{i=1}^n 2^{-i} 2^{i-2} = \Omega(n)$. Now, consider an alternative schedule that probes node i in each step with probability $2^{-i/2}/Z$, where $Z = \sum_{j=1}^n 2^{-j/2}$. The expected number of steps between two probes of i is $Z/2^{-i/2}$, and the cost of this schedule is

$$\sum_{i=1}^n 2^{-i} \left(\frac{2^{-i/2}}{\sum_{j=1}^n 2^{-j/2}} \right)^{-1} = \left(\sum_{j=1}^n 2^{-j/2} \right)^2 = O(1).$$

Thus, optimizing immediate gain is not optimal in this problem.

4.2 Lower Bound on Optimal Cost

In this section we provide a lower bound on the optimal cost, i.e., the cost of an optimal schedule.

Theorem 1. *For any c -schedule \mathcal{O} with finite cost we have*

$$\text{cost}(\mathcal{O}, \pi) \geq \max \left\{ \sum_{i=1}^n \pi_i, \frac{1}{2c} \left(\sum_{i=1}^n \sqrt{\pi_i} \right)^2 \right\}.$$

Proof. First, by Corollary 2, $\text{cost}(\mathcal{O}, \pi) \geq \sum_{i=1}^n \pi_i$. Now we show $\text{cost}(\mathcal{O}, \pi) \geq \frac{1}{2c} \left(\sum_{i=1}^n \sqrt{\pi_i} \right)^2$. Fix a positive integer $t > 0$, and suppose during the time interval $[0, t]$, \mathcal{O} probes node i at steps t_1, t_2, \dots, t_{n_i} . Let $t_0 = 0$ and $t_{n_i+1} = t$. So, the sequence t_0, \dots, t_{n_i+1} partition the interval $[0, t]$ into $n_i + 1$ intervals

$I_i(j) = [t_j + 1, t_{j+1}]$, for $0 \leq j \leq n_i$, and the length of $I_i(j)$ is $\ell_i(j) = t_{j+1} - t_j$. Applying the Cauchy-Schwartz inequality we have:

$$\begin{aligned} \sum_{j=0}^{n_i} \ell_i(j)^2 \sum_{j=0}^{n_i} 1 &\geq \left(\sum_{j=0}^{n_i} \ell_i(j) \right)^2 \\ \Rightarrow \sum_{j=0}^{n_i} \ell_i(j)^2 &\geq \frac{1}{n_i + 1} \left(\sum_{j=0}^{n_i} \ell_i(j) \right)^2 = \frac{t^2}{n_i + 1} = \frac{t^2}{n_i} \left(1 - \frac{1}{n_i + 1} \right). \end{aligned}$$

For $t' \in I_i(j)$, $Q_i^{\mathcal{O}}(t')$ is a Poisson random variable with parameter $\pi_i(t' - t_j)$. Therefore,

$$\begin{aligned} \sum_{t=1}^t \mathbb{E}[Q_i^{\mathcal{O}}(t')] &= \sum_{j=0}^{n_i} \sum_{t' \in I_i(j)} \mathbb{E}[Q_i^{\mathcal{O}}(t')] = \pi_i \sum_{j=0}^{n_i} (1 + \dots + \ell_i(j)) \\ &= \pi_i \sum_{j=0}^{n_i} \frac{\ell_i(j)(\ell_i(j) + 1)}{2} \geq \frac{\pi_i}{2} \sum_{j=0}^{n_i} \ell_i(j)^2 \geq \frac{\pi_i}{2} \frac{t^2}{n_i} \left(1 - \frac{1}{n_i + 1} \right). \end{aligned}$$

By summing over all nodes and averaging over t , we have

$$\begin{aligned} \sum_{i=1}^n \sum_{t'=1}^t \frac{1}{t} \mathbb{E}[Q_i^{\mathcal{O}}(t')] &\geq \sum_{i=1}^n \frac{1}{t} \frac{\pi_i}{2} \frac{t^2}{n_i} \left(1 - \frac{1}{n_i + 1} \right) \\ &= \sum_{i=1}^n \frac{\pi_i}{2} \frac{t}{n_i} \left(1 - \frac{1}{n_i + 1} \right) \geq \frac{1}{c} \left(\sum_{i=1}^n \frac{n_i}{t} \right) \left(\sum_{i=1}^n \frac{\pi_i}{2} \frac{t}{n_i} \left(1 - \frac{1}{n_i + 1} \right) \right) \\ &\geq \frac{1}{2c} \left(\sum_{i=1}^n \sqrt{\pi_i} \sqrt{\left(1 - \frac{1}{n_i + 1} \right)} \right)^2, \end{aligned} \tag{3}$$

where in the second line we use the fact that if the schedule executed c probes in each step then $\sum_{i=1}^n \frac{n_i}{t} \leq c$, and the third line is obtained by applying the Cauchy-Schwartz inequality.

It remains to show that for any schedule with finite cost, and any i such that $\pi_i > 0$, $\lim_{t \rightarrow \infty} n_i = \infty$. For sake of contradiction assume that there is a time s such that the node i is never probed by \mathcal{O} at time $t > s$. So, $\mathbb{E}[Q_i^{\mathcal{O}}(t)] = \pi(t - s)$ and we have $\text{cost}(\mathcal{O}, \pi) \geq \frac{1}{t} \sum_{t'=s}^t \mathbb{E}[Q_i^{\mathcal{O}}](t) = \frac{\pi_i}{t} \frac{(t-s)(t-s-1)}{2}$ which converges to ∞ as $t \rightarrow \infty$, which is a contradiction. Hence, for all i , $\lim_{t \rightarrow \infty} n_i = \infty$, and using (4.2) we obtain

$$\text{cost}(\mathcal{O}, \pi) \geq \lim_{t \rightarrow \infty} \frac{1}{2c} \left(\sum_{i=1}^n \sqrt{\pi_i} \sqrt{\left(1 - \frac{1}{n_i + 1} \right)} \right)^2 = \frac{1}{2c} \left(\sum_{i=1}^n \sqrt{\pi_i} \right)^2,$$

which completes the proof. \square

4.3 Deterministic $(3 + (c - 1)/c)$ -Approximation Schedule

We construct a deterministic 1-schedule in which each node i is probed approximately every $n_i = \frac{\sum_{j=1}^n \sqrt{\pi_j}}{\sqrt{\pi_i}}$ steps, and using that, present our $(3 + (c - 1)/c)$ -approximation schedule. For each i let r_i be a nonnegative integer such that $2^{r_i} \geq n_i > 2^{r_i-1}$, and let $\rho = \max_i r_i$.

Lemma 2. *There is a 2^ρ -cyclic 1-schedule \mathcal{D} such that node i is probed exactly every 2^{r_i} steps.*

Proof. Without loss of generality assume $\sum_{i=1}^n 2^{-r_i} = 1$, otherwise we can add auxiliary nodes to complete the sum to 1, with the powers (r_i 's) associated with the auxiliary nodes all bounded by ρ .

We prove the lemma by induction on ρ . If $\rho = 0$, then there is only one node, and the schedule is 1-cyclic. Now, assume the statement holds for all $\rho' < \rho$. Since the smallest frequency is $2^{-\rho}$, and the sum of the frequencies is 1, there must be two nodes, v and u , with same frequency $2^{-\rho}$. Join the two nodes to a new node w with frequency $2^{-\rho+1}$. Repeat this process for all nodes with frequency $2^{-\rho}$. We are left with a collection of nodes all with frequencies $> 2^{-\rho}$. By the inductive hypothesis there is a $(2^{\rho-1})$ -cyclic schedule \mathcal{D}' such that each node i is probed exactly every 2^{r_i} steps. In particular a node w that replaced u and v is probed exactly every $2^{-\rho+1}$ steps.

Now, we create an 2^ρ -schedule, \mathcal{D} , whose cycle is obtained by repeating the cycle of \mathcal{D}' two times. For each probe to w that replaced a pair u, v , in the first cycle we probe u and in the second cycle we probe v . Thus, u and v are probed exactly every 2^ρ steps, and the new schedule does not change the frequency of probing nodes with frequency larger than $2^{-\rho}$. \square

Theorem 2. *The cost of the deterministic 1-schedule \mathcal{D} is no more than 3 times of the optimal cost.*

Proof. By Lemma 2 each node i is probed exactly every 2^{r_i} steps. Using $2^{r_i-1} < \frac{\sum_{j=1}^n \sqrt{\pi_j}}{\sqrt{\pi_i}}$ we have $2^{r_i} + 1 \leq \frac{2 \cdot \sum_{j=1}^n \sqrt{\pi_j}}{\sqrt{\pi_i}} + 1$, and therefore

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{t'=1}^t \mathbb{E}[Q_i^{\mathcal{D}}(t')] &= \lim_{t \rightarrow \infty} \frac{1}{t} \frac{t}{2^{r_i}} \sum_{t'=1}^{2^{r_i}} \mathbb{E}[Q_i^{\mathcal{D}}(t')] = \frac{1}{2^{r_i}} \sum_{t'=1}^{2^{r_i}} \pi_i t' = \frac{\pi_i}{2^{r_i}} \frac{2^{r_i}(2^{r_i} + 1)}{2} \\ &\leq \frac{\pi_i}{2} \left(\frac{2 \sum_{j=1}^n \sqrt{\pi_j}}{\sqrt{\pi_i}} + 1 \right) = \sqrt{\pi_i} \cdot \sum_{j=1}^n \sqrt{\pi_j} + \frac{\pi_i}{2}. \end{aligned}$$

Thus by Theorem 1, we have

$$\begin{aligned} \text{cost}(\mathcal{D}, \pi) &= \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^n \sum_{t'=1}^t \mathbb{E}[Q_i^{\mathcal{D}}(t')] \leq \sum_{i=1}^n \left(\sqrt{\pi_i} \cdot \sum_{j=1}^n \sqrt{\pi_j} \right) + \frac{1}{2} \sum_{i=1}^n \pi_i \\ &= \left(\sum_{j=1}^n \sqrt{\pi_j} \right)^2 + \frac{1}{2} \sum_{j=1}^n \pi_j \leq 3 \cdot \text{cost}(\mathcal{O}, \pi), \end{aligned}$$

where $\text{cost}(\mathcal{O}, \pi)$ is the optimal cost. \square

Using the previous deterministic 1-schedule, the following corollary provides a c -schedule whose cost is within $(3 + (c - 1)/c)$ factor of the optimal cost.

Corollary 3. *There is a deterministic c -schedule \mathcal{D}^c whose cost is at most $(3 + (c - 1)/c)$ times of the optimal cost.*

Proof. Consider the execution of the deterministic 1-schedule \mathcal{D} constructed in Theorem 2 on generating vector $\frac{1}{c}\pi$. Let \mathcal{D}^c be a deterministic c -schedule obtained by grouping c consecutive probes of \mathcal{D} into one step. Suppose \mathcal{O} is an optimal c -schedule. Applying equation (1),

$$\begin{aligned} \text{cost}(\mathcal{D}^c, \pi) &= \sum_{i=1}^n \pi_i \omega_i^{\mathcal{D}^c} = \sum_{i=1}^n \frac{\pi_i}{c} c \omega_i^{\mathcal{D}^c} \leq \sum_{i=1}^n \frac{\pi_i}{c} (\omega_i^{\mathcal{D}} + c - 1) \\ &= \text{cost}(\mathcal{D}, \pi) + \sum_{i=1}^n \frac{(c-1)\pi_i}{c} \leq 3 \left(\sum_{i=1}^n \sqrt{\frac{\pi_i}{c}} \right)^2 + \sum_{i=1}^n \frac{(c-1)\pi_i}{c} \\ &= \frac{3}{2c} \left(\sum_{i=1}^n \sqrt{\pi_i} \right)^2 + \sum_{i=1}^n \frac{(c-1)\pi_i}{c} \leq \left(3 + \frac{c-1}{c} \right) \text{cost}(\mathcal{O}, \pi), \end{aligned}$$

where the first inequality holds because some items could be detected in less than c steps in the 1-schedule but are counted in one full step of the c -schedule. The last inequality is obtained by applying Theorem 1. \square

4.4 On Optimal Memoryless Schedule

Here, we consider memoryless schedules, and show that the memoryless 1-schedule with minimum cost can be easily computed. We call a memoryless schedule with minimum cost among memoryless schedules, an optimal memoryless schedule. We also provide an upper bound on the minimum cost of a memoryless c -schedule.

Theorem 3. *Let $\mathcal{R} = (p_1, \dots, p_n)$ be a memoryless 1-schedule. Then $\text{cost}(\mathcal{R}, \pi) \geq \left(\sum_{i=1}^n \sqrt{\pi_i} \right)^2$, and the equality holds if and only if $p_i = \frac{\sqrt{\pi_i}}{\sum_{j=1}^n \sqrt{\pi_j}}$, for all i .*

Proof. Since probing each node i is a geometric distribution with parameter p_i , the expected time until an item generated in node i is discovered, is $\omega_i^{\mathcal{R}} = 1/p_i$. Therefore, by Lemma 1, we have $\text{cost}(\mathcal{R}, \pi) = \sum_{i=1}^n \frac{\pi_i}{p_i}$. We find $p^* = \arg\min_{\mathcal{S}=p} \text{cost}(\mathcal{R}, \pi)$, using the Lagrange multipliers:

$$\frac{\partial}{\partial p_j} \left(\sum_{i=1}^n \frac{\pi_i}{p_i} + \lambda \sum_{i=1}^n p_i \right) = 0 \implies p_j \propto \sqrt{\pi_j}.$$

Therefore, $\text{cost}(\mathcal{R}, \pi)$ is minimized if $p_i = \frac{\sqrt{\pi_i}}{\sum_{j=1}^n \sqrt{\pi_j}}$, and in this case the (minimized) cost will be

$$\text{cost}(\mathcal{R}, \pi) = \sum_{i=1}^n \left(\sqrt{\pi_i} \cdot \sum_{j=1}^n \sqrt{\pi_j} \right) = \left(\sum_{i=1}^n \sqrt{\pi_i} \right)^2. \quad \square$$

Corollary 4. *The cost of the optimal memoryless 1-schedule is within a factor of 2 of the cost of any optimal 1-schedule.*

Proof. The cost of the schedule \mathcal{R} in Theorem 3 is $\left(\sum_{i=1}^n \sqrt{\pi_i} \right)^2$, which is bounded by $2 \cdot \text{cost}(\mathcal{O}, \pi)$ for an optimal 1-schedule \mathcal{O} using Theorem 1. \square

Corollary 5. *There is memoryless c -schedule, \mathcal{R}^c , whose cost is within a factor of $(2 + (c - 1)/c)$ of any optimal c -schedule.*

Proof. Suppose \mathcal{R}^c is a memoryless c -schedule obtained by choosing c probes in each step, each chosen according to the optimal memoryless 1-schedule, \mathcal{R} , computed in Theorem 3. Using the same argument as in the proof of Corollary 3 we have

$$\text{cost}(\mathcal{R}^c, \pi) \leq \frac{1}{c} \left(\sum_{i=1}^n \sqrt{\pi_i} \right)^2 + \frac{c-1}{c} \sum_{i=1}^n \pi_i \leq \left(2 + \frac{c-1}{c} \right) \text{cost}(\mathcal{O}, \pi),$$

for an optimal c -schedule \mathcal{O} . \square

4.5 On Adaptive Algorithm for Memoryless Schedules

Assume now that the scheduling algorithm starts with no information on the generating vector π (or that the vector has changed). We design and analyze an adaptive algorithm, **Adaptive**, that outputs a schedule \mathcal{A} convergent to the optimal memoryless algorithm \mathcal{R} (see Section 4.4) by gradually learning the vector π by observing the system. To simplify the presentation we present and analyze a 1-schedule algorithm. The results easily scale up to any integer $c > 1$, where the adaptive algorithm outputs a c -schedule convergent to \mathcal{R}^c (as in Section 4.4).

Each iteration of the algorithm **Adaptive** starts with an estimate $\tilde{\pi} = (\tilde{\pi}_1, \dots, \tilde{\pi}_n)$ of the unknown generating vector $\pi = (\pi_1, \dots, \pi_n)$. Based on this estimate the algorithm chooses to probe node i with probability $p_i(t) = \frac{\sqrt{\tilde{\pi}_i}}{\sum_{j=1}^n \sqrt{\tilde{\pi}_j}}$ (which is the optimal memoryless schedule if $t\pi$ was the correct estimate). If nodes i is

probed at time t , the estimate of π_i is updated to $\tilde{\pi}_{i_0} \leftarrow \frac{\max(1, c_{i_0})}{t}$, where c_{i_0} is the total number of new items discovered in that node since time 0.

Algorithm 1: Adaptive

Outputs: $\mathcal{A}(t)$, for $t = 1, 2, \dots$
begin
 $(c_1, \dots, c_n) \leftarrow (0, \dots, 0);$
 $(\tilde{\pi}_1, \dots, \tilde{\pi}_n) \leftarrow (1, \dots, 1);$
 for $t = 1, 2, \dots$ **do**
 for $i \in \{1, \dots, n\}$ **do**
 $p_i(t) \leftarrow \frac{\sqrt{\tilde{\pi}_i}}{\sum_{j=1}^n \sqrt{\tilde{\pi}_j}};$
 $\mathcal{A}(t) \sim p(t);$
 output $\mathcal{A}(t);$
 $c' \leftarrow$ number of new items caught at i_0 ;
 $c_{i_0} \leftarrow c_{i_0} + c';$
 $\tilde{\pi}_{i_0} \leftarrow \frac{\max(1, c_{i_0})}{t};$

We denote the output of **Adaptive** schedule by \mathcal{A} and the optimal memoryless 1-schedule by $\mathcal{R} = p^* = (p_1^*, \dots, p_n^*)$; see Section 4.4. Our main result of this section is the following theorem.

Theorem 4. *The schedule \mathcal{A} converges to \mathcal{R} , and thus, $\text{cost}(\mathcal{A}, \pi) = \text{cost}(\mathcal{R}, \pi)$.*

To prove Theorem 4 we need the following lemmas.

Lemma 3. *For any time t and $i \in [n]$ we have $p_i(t) \geq \frac{1}{n\sqrt{t}}$.*

Proof. It is easy to see that $p_i(t)$ will reach its lowest value at time t only if for $j \neq i$ we have $\tilde{\pi}_j = 1$ and $\tilde{\pi}_i = \frac{1}{t-1}$ (which requires i to be probed at time $t-1$). Therefore, $p_i(t) \geq \frac{1/\sqrt{t-1}}{1/\sqrt{t-1} + n-1} = \frac{1}{1+(n-1)\sqrt{t-1}} \geq \frac{1}{n\sqrt{t}}$. \square

Define $\delta(t) = 4n \exp\left(-\frac{\pi_* t^{1/3}}{6}\right)$ and let N_0 be the smallest integer t such that $\exp\left(-\frac{\sqrt{t}}{2n}\right) \leq 2 \exp\left(-\frac{\pi_* t^{1/3}}{6}\right)$. Note that one can choose $\delta(t) = 4n \exp\left(-\frac{\pi_* t^{1/2-\epsilon}}{6}\right)$ for any $\epsilon \in (0, 1/2)$, and for convenience we chose $\epsilon = 1/6$.

Lemma 4. *For any time $t \geq N_0$, with probability $\geq 1 - \delta(t)/2$, all the nodes are probed during the time interval $[t/2, t]$.*

Proof. By Lemma 3, the probability of not probing i during the time interval $[t/2, t]$ is at most

$$\prod_{t'=t/2}^{t-1} (1 - p_i(t')) \leq \left(1 - \frac{1}{n\sqrt{t}}\right)^{t/2} \leq e^{-\frac{t}{2n\sqrt{t}}} \leq 2 \exp\left(-\frac{\pi_* t^{1/3}}{6}\right) = \frac{\delta(t)}{2n}.$$

A union bound over all the nodes completes the proof. \square

Lemma 5. Suppose node i is probed at a time $t' > t/2$. Then,

$$\Pr \left[|\tilde{\pi}_i(t') - \pi_i| > t^{-\frac{1}{3}} \pi_i \right] < \frac{\delta(t)}{2n}.$$

Proof. We estimate π from $t' > t/2$ steps, each with π_i expected number of new items. Applying a Chernoff bound [18] for the sum of t' independent random variables with either Bernulli or Poisson distribution we have

$$\Pr \left[|\tilde{\pi}_i(t') - \pi_i| > t^{-\frac{1}{3}} \pi_i \right] < 2 \exp \left(-\frac{t^{-\frac{2}{3}} \pi_i t'}{3} \right) \leq 2 \exp \left(-\frac{t^{-\frac{2}{3}} \pi_i t}{6} \right) = \frac{\delta(t)}{2n}. \quad \square$$

Note that by union bound, Lemma 5 holds, with probability at least $1 - \delta(t)/2$, for all the nodes that are probed after $t/2$.

Lemma 6. Suppose $t \geq N_0$. With probability at least $1 - \delta(t)$ we have for all $i \in [n]$,

$$\left(1 - \frac{1}{t^{1/3} + 1}\right) p_i^* \leq \sqrt{\frac{1 - t^{-1/3}}{1 + t^{-1/3}}} p_i^* \leq p_i(t) \leq \sqrt{\frac{1 + t^{-1/3}}{1 - t^{-1/3}}} p_i^* \leq \left(1 + \frac{1}{t^{1/3} - 1}\right) p_i^*$$

Proof. Applying Lemma 4, Lemma 5 and a union bound, with probability $1 - \delta(t)$ all the nodes are probed during the time $[t/2, t)$ and $|\tilde{\pi}_i(t) - \pi_i| \leq t^{-1/3} \pi_i$ for all $i \in [n]$. Since $p_i(t) = \frac{\sqrt{\tilde{\pi}_i(t)}}{\sum_j \sqrt{\tilde{\pi}_j(t)}}$, we obtain

$$p_i(t) \geq \frac{\sqrt{(1 - t^{-1/3}) \pi_i}}{\sum_j \sqrt{(1 + t^{-1/3}) \pi_j}} = \sqrt{\frac{1 - t^{-1/3}}{1 + t^{-1/3}}} \frac{\sqrt{\pi_i}}{\sum_j \sqrt{\pi_j}} = \sqrt{\frac{1 - t^{-1/3}}{1 + t^{-1/3}}} p_i^* \geq \left(1 - \frac{1}{t^{1/3} + 1}\right) p_i^*$$

where the last inequality uses the Taylor series of $\sqrt{1 + x}$. The upper bound is obtained by a similar argument. \square

Corollary 6. The variation distance between the distribution $p(t) = (p_1(t), \dots, p_n(t))$ used by algorithm *Adaptive* at time $t \geq N_0$, and the distribution $p^* = (p_1^*, \dots, p_n^*)$ used by the optimal memoryless algorithm satisfy

$$\|p(t) - p^*\| = \frac{1}{2} \sum_{i=1}^n |p_i(t) - p_i^*| \leq \frac{n}{t^{1/3} - 1} + \delta(t) \xrightarrow{t \rightarrow \infty} 0.$$

Finally, we present our proof for Theorem 4.

Proof of Theorem 4. Recall that we defined $\tau_i^S(t)$ as the number of steps from the last time that node i was probed until time t in an execution of an schedule \mathcal{S} , and $\mathbb{E}[Q_i^S] = \pi_i \mathbb{E}[\tau_i^S(t)]$.

Let $F(t)$ indicate the event that the inequalities in Lemma 6 are held for $\forall t' \in [t/2, t)$. Therefore, $\Pr[F(t)] < 1 - \frac{\delta(t/2) \cdot t}{2}$ by applying union bound over all $t' \in [t/2, t)$, and using the fact that $\delta(t') \leq \delta(t/2)$. Therefore,

$$\begin{aligned} |\mathbb{E}[Q^{\mathcal{A}}(t)] - \mathbb{E}[Q^{\mathcal{R}}(t)]| &= \left| \sum_{i=1}^n \pi_i \mathbb{E}[\tau_i^{\mathcal{A}}(t)] - \sum_{i=1}^n \pi_i \mathbb{E}[\tau_i^{\mathcal{R}}(t)] \right| \\ &\leq \sum_{i=1}^n \pi_i |\mathbb{E}[\tau_i^{\mathcal{A}}(t)] - \mathbb{E}[\tau_i^{\mathcal{R}}(t)]| \\ &\leq \sum_{i=1}^n \pi_i |\mathbb{E}[\tau_i^{\mathcal{A}}(t)] - \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)]| \\ &\quad + \sum_{i=1}^n \pi_i |\mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] - \mathbb{E}[\tau_i^{\mathcal{R}}(t)]|, \end{aligned}$$

where we used the triangle inequality for both inequalities. So, it suffices to show that for every i ,

$$\lim_{t \rightarrow \infty} |\mathbb{E}[\tau_i^{\mathcal{A}}(t)] - \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)]| = \lim_{t \rightarrow \infty} |\mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] - \mathbb{E}[\tau_i^{\mathcal{R}}(t)]| = 0.$$

Obviously, $\tau_i^{\mathcal{A}}(t) \leq t$. Now by letting $t \geq 2N_0$ we have,

$$\begin{aligned} \mathbb{E}[\tau_i^{\mathcal{A}}(t)] &= \Pr[F(t)] \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] + \Pr[\neg F(t)] \mathbb{E}[\tau_i^{\mathcal{A}}(t) | \neg F(t)] \\ &\leq \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] + \frac{\delta(t/2)t}{2} = \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] + \frac{\delta(t/2)t^2}{2}. \end{aligned} \quad (4)$$

We also get

$$\begin{aligned} \mathbb{E}[\tau_i^{\mathcal{A}}(t)] &\geq \left(1 - \frac{\delta(t/2)t}{2}\right) \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] \\ &= \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] - \frac{\delta(t/2)t}{2} \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] \geq \mathbb{E}[\tau_i^{\mathcal{A}}(t) | X] - \frac{\delta(t)t^2}{2}. \end{aligned} \quad (5)$$

Note that $\lim_{t \rightarrow \infty} \frac{\delta(t/2)t^2}{2} = \lim_{t \rightarrow \infty} 4ne^{-\frac{\pi_* t^{1/3}}{6\sqrt[3]{2}}} t^2 = 0$, and thus by (4) and (5) we have

$$\lim_{t \rightarrow \infty} \mathbb{E}[\tau_i^{\mathcal{A}}(t)] - \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] = 0 \Rightarrow \lim_{t \rightarrow \infty} |\mathbb{E}[\tau_i^{\mathcal{A}}(t)] - \mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)]| = 0. \quad (6)$$

Now, we show that $\lim_{t \rightarrow \infty} |\mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] - \mathbb{E}[\tau_i^{\mathcal{R}}(t)]| = 0$. So here, we assume $F(t)$ holds. So for every $i \in [n]$, node i is probed in $[t/2, t)$, and for all $t' \in [t/2, t)$ we have

$$(i) \ p_i(t') \geq \left(1 - \frac{1}{t'^{1/3}+1}\right) p_i^* \geq \left(1 - \frac{1}{(t/2)^{1/3}+1}\right) p_i^*.$$

$$\mathbb{E}[\tau_i^{\mathcal{A}}(t) | F(t)] \leq \left(1 - \frac{1}{(t/2)^{1/3}+1}\right)^{-1} \frac{1}{p_i^*} = \left(1 + (t/2)^{-1/3}\right) \frac{1}{p_i^*}.$$

(ii) $p_i(t') \leq \left(1 + \frac{1}{t^{1/3}-1}\right) p_i^* \leq \left(1 + \frac{1}{(t/2)^{1/3}-1}\right) p_i^*$. Hence,

$$\mathbb{E} [\tau_i^{\mathcal{A}}(t) \mid F(t)] \geq \left(1 + \frac{1}{(t/2)^{1/3}-1}\right)^{-1} \frac{1}{p_i^*} = \left(1 - (t/2)^{-1/3}\right) \frac{1}{p_i^*}.$$

Obviously $\mathbb{E} [\tau_i^{\mathcal{R}}(t)] = \frac{1}{p_i^*}$, since probing node i by \mathcal{R} can be viewed as a geometric distribution with parameter p_i^* , and since $\delta(t) \rightarrow 0$ as $t \rightarrow \infty$ we have

$$\begin{aligned} \mathbb{E} [\tau_i^{\mathcal{R}}(t)] &= \frac{1}{p_i^*} = \lim_{t \rightarrow \infty} \left(1 - (t/2)^{-1/3}\right) \frac{1}{p_i^*} \leq \lim_{t \rightarrow \infty} \mathbb{E} [\tau_i^{\mathcal{A}}(t) \mid F(t)] \\ &\leq \lim_{t \rightarrow \infty} \left(1 + (t/2)^{-1/3}\right) \frac{1}{p_i^*} = \frac{1}{p_i^*} = \mathbb{E} [\tau_i^{\mathcal{R}}(t)]. \end{aligned}$$

Therefore,

$$\lim_{t \rightarrow \infty} |\mathbb{E} [\tau_i^{\mathcal{A}}(t) \mid F(t)] - \mathbb{E} [\tau_i^{\mathcal{R}}(t)]| = 0. \quad (7)$$

Thus, by (6) and (7) we have $\lim_{t \rightarrow \infty} |\mathbb{E} [Q^{\mathcal{A}}(t)] - \mathbb{E} [Q^{\mathcal{R}}(t)]| = 0$, and \mathcal{A} converges to \mathcal{S} , and since $\lim_{t \rightarrow \infty} \mathbb{E} [Q^{\mathcal{S}}(t)] = (\sum_{i=1}^n \sqrt{\pi_i})^2$, it implies that $\lim_{t \rightarrow \infty} \mathbb{E} [Q^{\mathcal{A}}(t)] = (\sum_{i=1}^n \sqrt{\pi_i})^2 = \text{cost}(\mathcal{A}, \pi)$ (by Cesaro Mean [6]). \square

Note that one can obtain an adaptive schedule \mathcal{A}^c by choosing c probes in each step, at each round of **Adaptive**, and using similar argument as in Section 4.4 (and similar to Corollary 3), it is easy to see that \mathcal{A}^c converges to \mathcal{R}^c .

Finally, if π changes, the **Adaptive** algorithm converges to the new optimal memoryless algorithm, as the change in the rate of generating new items is observed by **Adaptive**.

References

1. Bright, L., Gal, A., Raschid, L.: Adaptive pull-based policies for wide area data delivery. *ACM Transactions on Database Systems (TODS)* 31(2), 631–671 (2006)
2. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: *Proceedings of the Tenth International Workshop on Multimedia Data Mining*. pp. 4:1–4:10. MDMKDD '10, ACM, New York, NY, USA (2010)
3. Dasgupta, A., Ghosh, A., Kumar, R., Olston, C., Pandey, S., Tomkins, A.: The discoverability of the web. In: *Proceedings of the 16th international conference on World Wide Web*. pp. 421–430. ACM (2007)
4. Delaney, A.: The growing role of news in trading automation (Oct 2009), http://www.machinereadablenews.com/images/dl/Machine_Readable_News_and_Algorithmic_Trading.pdf
5. Group, D.B.: Alphaflash trader automated trading based on economic events (2015), <http://www.alphaflash.com/product-info/alphaflash-trader>
6. Hardy, G.H.: *Divergent series*, vol. 334. American Mathematical Soc. (1991)

7. Hart, W., Murray, R.: Review of sensor placement strategies for contamination warning systems in drinking water distribution systems. *Journal of Water Resources Planning and Management* 136(6), 611–619 (2010)
8. Hope, B.: How computers trawl a sea of data for stock picks. *The Wall Street Journal* (Apr 2015), <http://www.wsj.com/articles/how-computers-trawl-a-sea-of-data-for-stock-picks-1427941801?KEYWORDS>
9. Horincar, R., Amann, B., Artières, T.: Online refresh strategies for content based feed aggregation. *World Wide Web* pp. 1–35 (2014)
10. Krause, A., Leskovec, J., Guestrin, C., VanBriesen, J., Faloutsos, C.: Efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management* 134(6), 516–526 (2008)
11. Latar, N.L.: The robot journalist in the age of social physics: The end of human journalism? In: *The New World of Transitioned Media*, pp. 65–80. Springer (2015)
12. Leon-Garcia, A.: *Probability, Statistics, and Random Processes For Electrical Engineering* (3rd Edition). Prentice Hall, 3 edn. (1 2008)
13. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J.M., Glance, N.S.: Cost-effective outbreak detection in networks. In: Berkhin, P., Caruana, R., Wu, X. (eds.) *KDD*. pp. 420–429. ACM (2007)
14. Ltd., E.A.: Discovering the web’s hidden alpha (Jun 2014), http://www.eaglealpha.com/whitepaper_pdf
15. Mathioudakis, M., Koudas, N.: Twittermonitor: Trend detection over the twitter stream. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. pp. 1155–1158. SIGMOD ’10, ACM, New York, NY, USA (2010)
16. McKinney, W.: *Structured Data Challenges in Finance and Statistics* (Nov 2011), <http://www.slideshare.net/wesm/structured-data-challenges-in-finance-and-statistics>
17. Mitra, G., Mitra, L.: *The handbook of news analytics in finance*, vol. 596. John Wiley & Sons (2011)
18. Mitzenmacher, M., Upfal, E.: *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press (2005)
19. Oita, M., Senellart, P.: Deriving dynamics of web pages: A survey. In: *TWAW (Temporal Workshop on Web Archiving)* (2011)
20. Ostfeld, A., et al.: The battle of the water sensor networks (bwsn): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management* 134(6), 556–568 (2008)
21. Sia, K.C., Cho, J., Cho, H.K.: Efficient monitoring algorithm for fast news alerts. *Knowledge and Data Engineering, IEEE Transactions on* 19(7), 950–961 (2007)
22. Wolf, J.L., Squillante, M.S., Yu, P., Sethuraman, J., Ozsen, L.: Optimal crawling strategies for web search engines. In: *Proceedings of the 11th international conference on World Wide Web*. pp. 136–147. ACM (2002)